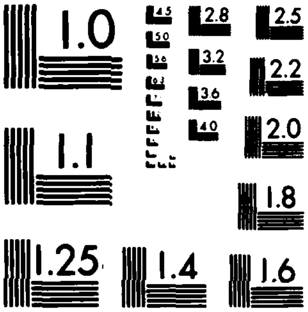
AD-A127 883   SOLUTION OF SYSTEMS OF COMPLEX LINEAR EQUATIONS IN THE   1/1
L INFINITY NORM WI..(U) NAVAL UNDERWATER SYSTEMS CENTER
NEW LONDON CT NEW LONDON LAB.. R L STREIT 18 APR 83
UNCLASSIFIED   NOSC-TR-6913   F/G 12/2   NL

END
DATE
FILMED
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

1

NUSC Technical Report 6913
18 April 1983

Solution of Systems of Complex Linear Equations in the L_∞ Norm With Constraints on the Unknowns

Roy L. Streit
Information Services Department

Naval Underwater Systems Center
Newport, Rhode Island / New London, Connecticut

Approved for public release; distribution unlimited.

Preface

This work was supported jointly by the Office of Naval Research (Project RR014-07-01, Code 434, Dr. Neal Glassman) and by the Independent Research Program of the Naval Underwater Systems Center, IR/IED Project No. A70210.

The Technical Reviewer for this report was Dr. A. H. Nuttall, Code 3302.

Reviewed and Approved: 18 April 1983

W. A. Von Winkle
Associate Technical Director for Technology

The author of this report is located at the
Naval Underwater Systems Center, New London
Laboratory, New London, Connecticut 06320.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TR 6913	2. GOVT ACCESSION NO. AD-A127883	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) SOLUTION OF SYSTEMS OF COMPLEX LINEAR EQUATIONS IN THE L_∞ NORM WITH CONSTRAINTS ON THE UNKNOWNS		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Roy L. Streit		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Underwater Systems Center New London Laboratory New London, Connecticut 06320		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Underwater Systems Center Newport, RI 02841		12. REPORT DATE 18 April 1983 13. NUMBER OF PAGES 55
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

complex linear equations	semi-infinite programming
Chebyshev solution	complex approximation
convex constraints	L_∞ norm

infinity

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An algorithm for the numerical solution of general systems of complex linear equations in the L_∞, or Chebyshev, norm is presented. The objective is to find complex values for the unknowns so that the maximum magnitude residual of the system is a minimum. The unknowns are required to satisfy certain general convex constraints. In particular, bounds on the magnitudes of the unknowns are imposed. In the algorithm presented here, this problem is replaced by a linearized problem. The linearized problem is a linear

20. (Cont'd)

program which is generated in such a way that the relative error between its (exact) solution and the (exact) solution of the original problem can be estimated without knowing a priori the solution of either. Furthermore, the maximum relative error can easily be made as small as desired by selecting an appropriate linearized problem. Order of magnitude improvements in both computation time and computer storage requirements in an implementation of the simplex algorithm to this linear program are presented. Three numerical examples are included, one of which is a discretized complex function approximation problem. Extrapolation and an active set method are suggested for solving the general problem when greater accuracy is required than can be obtained economically by solving the linearized problems.

TABLE OF CONTENTS

		Page
I.	Introduction	1
II.	Solution of the linearized problem	10
III.	Solution of the linearized problem for large p.	28
IV.	Details of the revised simplex algorithm	32
V.	Examples	36
VI.	Concluding remarks	51
	Bibliography	54

SOLUTION OF SYSTEMS OF COMPLEX LINEAR EQUATIONS IN THE L_∞ NORM WITH CONSTRAINTS ON THE UNKNOWNS

I. Introduction

The numerical solution of general systems of complex linear equations in the L_∞, or Chebyshev, norm is a mathematical problem that arises in several applications. The objective is to find complex values for all the unknowns so that the maximum magnitude residual of the system of equations is a minimum. The unknowns are not allowed to assume any complex value whatever; instead, they are required to satisfy convex constraints of the form that can occur in the applications.

To be specific, let n, m, and ℓ be integers satisfying $m \geq 1$, $n \geq 1$, and $\ell \geq 1$. Let the matrices $A \in C^{n \times m}$, $B \in C^{n \times \ell}$, and the row vectors $f \in C^m$, $g \in C^\ell$, $a \in C^n$, $d \in R^n$, and $c \in R^\ell$ be given. Let A_j and B_j denote the j-th columns of matrices A and B, respectively. The problem investigated in this report can be stated as a non-linear mathematical programming problem in the following manner.

<u>Problem</u>

$$\min_{\varepsilon \in R, z \in C^n} \varepsilon \tag{1}$$

subject to:

$$|zA_j - f_j| \leq \varepsilon \text{ , } j = 1,\ldots,m \tag{2}$$

$$|zB_j - g_j| \leq c_j, \; j = 1,\ldots,\ell \tag{3}$$

$$|z_j - a_j| \leq d_j, \; j = 1,\ldots,n \tag{4}$$

The vector of unknowns, z, is also taken to be a row vector. Notational convenience is the only reason for using row instead of column vectors. It is assumed throughout this report that $c_j > 0$ and $d_j > 0$ for all indices j. If any c_j or d_j were equal to 0, the corresponding inequality could be replaced by an equation involving no absolute values. These equations could then be used to produce a problem of the same mathematical form as (1) - (4), but having fewer variables. Further discussion of equality constraints is given later in this section.

If there are no constraints (3) - (4) and if the system is under-determined (i.e., $m \leq \text{rank } A \leq n$), then simpler methods than the one presented in this report suffice to find all the solutions of the problem. Adjoining the constraints (3) - (4), however, can give a theoretically interesting problem even if it is under-determined. For example, if $m = n = \text{rank } A$, then the problem (1) - (2) has a unique solution which can be found by solving the system $zA = f$. But if this solution does not satisfy the constraints (3) - (4), then it cannot be the solution of the problem (1) - (4). The method presented in this report can be applied whether the system is over-determined or under-determined with equal ease.

In the approach investigated here, this problem is replaced by a linearized problem. The linearized problem is a linear program which is generated in such a way that the relative error between its (exact) solution and the (exact) solution of the original non-linear problem can be estimated without knowing the solution of either. Furthermore, the maximum relative error can easily be made as small as desired by selecting an appropriate linearized problem. See Theorem 2 below.

The starting point for the linearized problem is the following simple fact. Let x and y denote real numbers and let i denote the imaginary unit. Then

$$|x + iy| = \max_{0 \leq \theta < 2\pi} (x \cos \theta + y \sin \theta) . \qquad (5)$$

A proof using the Cauchy-Schwartz inequality is trivial. The maximum in (5) is now approximated by using a finite number of angles in the interval $[0,2\pi)$. Let $p > 3$ be an integer, and let the set $\Theta = \{\theta_1,\ldots,\theta_p\}$ be given. Then we define the linearized absolute value

$$|x + iy|_\Theta \triangleq \max_{\theta\epsilon\Theta} \ (x \cos \theta + y \sin \theta). \tag{6}$$

The linearized version of the problem results from replacing every absolute value in (2) - (4) with linearized absolute values.

Linearized Problem

$$\min_{\epsilon\epsilon R, z\epsilon C^n} \ \epsilon \tag{7}$$

$$\text{subject to:} \quad |zA_j - f_j|_\Theta \leq \epsilon \ , \qquad j = 1,\ldots,m \tag{8}$$

$$|zB_j - g_j|_\Theta \leq c_j, \qquad j = 1,\ldots,\ell \tag{9}$$

$$|z_j - a_j|_\Theta \leq d_j, \qquad j = 1,\ldots,n \tag{10}$$

The same set Θ is used in the inequalities (8) - (10) merely for the sake of convenience; it is certainly possible to use a different Θ set for each of the $m+n+\ell$ inequalities (2) - (4). Substituting in real and imaginary parts for all complex quantities and applying the definition (6), it becomes clear that the linearized problem is a linear program in $2n + 1$ unknowns with $(m+n+\ell)p$ inequalities. See section II. Even for modest values of m,n,ℓ, and p, this linear program becomes very large quickly. For example, if $n = \ell = 20$, $m = 100$, and $p = 64$, it has 41 variables and 8960 inequalities. Furthermore, the inequalities (8) and (9) are 100% dense since, in the applications, the matrices A and B are normally 100% dense.

The linearized problem always has more inequalities than unknowns, so we solve it by solving its dual using the revised simplex method. The special structure of the dual of the linearized problem is used effectively in several ways. One prominent feature of our algorithm is a significant reduction in computer storage. Instead of $(m+n+\ell)(2n+1)p$ storage locations required in a straightforward application of the revised simplex method, the method presented here requires only $2n(m+\ell) + 2p$ storage locations. Another prominent feature of the algorithm is that the most negative reduced cost coefficient of all the variables in the dual problem is computed with only $O(mn \log_2 p)$ real multiplications, instead of the $O(mnp)$ real multiplications normally required. A third prominent feature of the algorithm is that the solution for large values of p is approached through smaller values of p. That is, we first solve the linearized problem for $p = 4$, and the solution for $p = 4$ is used as an advanced start for the linearized problem with $p = 8$. We continue in this fashion, doubling p at each step, until a specified value of p is attained. In this way, problems with p as large as 2048 have been successfully solved. These three features have been incorporated into a FORTRAN program which seems to be practical for modest values of m, n, and ℓ even for large values of p.

Equality constraints of the form $zH_j = e_j$, $j = 1,\dots,r$ with $1 \leq r \leq n$ can be added to the original problem if desired. One way to handle them in this context is by the method of elimination. If these equations have rank $q \leq r$, then some q of these unknowns can be solved for explicitly in terms of the remaining $n - q$ unknowns. These q unknowns can then be substituted out of the original problem, yielding a new problem with only $n - q$ unknowns. Moreover, the new problem has the same mathematical form as the original problem, so it may be solved in the manner proposed here. For this reason, equality constraints are not included in the original problem statement.

The set Θ used in (6) can be an arbitrary set of angles. However, it is convenient to restrict Θ in two ways. First, the number of points in Θ is required to be a multiple of 4 and a power of 2; that is, $p = 2^K$ with $K \geq 2$. Second, the points in Θ are required to be the arguments of the complex p-th roots of unity; that is,

$$\theta_k = (k-1)2\pi/p, \qquad k=1,2,\ldots, p. \tag{11}$$

These restrictions allow the doubling of p as mentioned above and simultaneously facilitate considerable computational efficiencies in determining the incoming basic variable. See section III below. Another reason to require (11) is that for fixed p it gives the tightest upper bound in the inequality (12) below. Unless specifically stated otherwise, Θ will be assumed to satisfy both these restrictions throughout this report.

There is one hazard in replacing the original problem (1) - (4) with the linearized problem (7) - (10). The linearized constraints have a larger feasible region than the original constraints, so it is possible that the linearized problem has solutions for some values of p even when the original problem is actually infeasible. The feasible region of the original problem is approximated more and more closely as p is increased, so the linearized problem must ultimately fail to have a solution for sufficiently large p when the original problem is infeasible. If the original problem is in some sense "nearly" feasible, but in reality is infeasible, the linearized problem may possess solutions for very large values of p. Thus one may be deceived in certain problems. An alternative viewpoint is that any false solution obtained in this manner to infeasible problems actually represents a "reasonable" solution to a poorly defined problem. Whether or not this view is sensible depends on the application. An example is given later in section V.

It is worth mentioning that the constraints (4) are merely instances of the constraints (3), and similarly for the constraints (10) and (9). Distinguishing these two kinds of constraints is desirable not only for the expected reasons of simplicity of form and economy of computer storage, but also because the simpler constraints present themselves implicitly in the unconstrained problem. That is, if the constraints (9) and (10) of the linearized problem are eliminated, and the dual of the resulting linear program solved by the simplex method, then some of the constraints (10) reappear in the guise of artificial variables. This will be made clear later in section II.

The following theorem lists some basic properties of the linearized absolute value. The inequalities (12) are especially useful.

Theorem 1. Let $p \geq 4$ be an even integer, and let the elements of the set Θ satisfy (11). Then, for $u \in C$ and $v \in C$,

i. $|u|_\Theta \geq 0$ and $|u|_\Theta = 0$ if and only if $u = 0$.

ii. $|u + v|_\Theta \leq |u|_\Theta + |v|_\Theta$.

iii. Given α complex, $|\alpha u|_\Theta = |\alpha| \cdot |u|_\Theta$ for all u if and only if $\arg u \in \Theta$.

iv. $|u|_\Theta = |-u|_\Theta$, $\lim_{\alpha_n \to 0} |\alpha_n u|_\Theta = 0$, and $\lim_{|u_n|_\Theta \to 0} |\alpha u_n| = 0$.

v. $|u|_\Theta \leq |u| \leq |u|_\Theta \sec(\pi/p)$. (12)

Proof. Properties (i) to (iv) are straightforward. The first inequality in (12) follows immediately from (5) and (6). The second inequality in (12) is easily proved by visualizing the set of all $u \in C$ such that $|u|_\Theta = 1$ as an equilateral polygon of p sides whose inscribed circle is $|u| = 1$. This concludes the proof. (Note that if the hypothesis is weakened to $p \geq 3$, the theorem remains valid except for the first equation in (iv).)

Although the linearized absolute value is not a norm on C, it is a quasi-norm because of properties (i), (ii), and (iv). See [1, pp. 30-32]. Also, the asymptotic result

$$\sec \frac{\pi}{p} = 1 + \frac{\pi^2}{2p^2}, \quad p \to \infty, \tag{13}$$

shows that to attain a relative accuracy of 5 significant digits (i.e., a relative error less than $.5 \times 10^{-5}$) in (12) requires that $p \geq 1024$.

The next theorem connects the solution of the original problem with the solution of the linearized problem.

Theorem 2. Let $\varepsilon'' \in R$ and $z'' \in C^n$ solve problem (1) - (4), and let $\varepsilon' \in R$ and $z' \in C^n$ solve the linearized problem (7) - (10). Then

$$\varepsilon' \leq \varepsilon'' \leq \varepsilon' \sec(\pi/p) \tag{14}$$

and

$$|z'A_j - f_j| \leq \varepsilon' \sec(\pi/p) \ , \ j = 1,\ldots,m \tag{15}$$

$$|z'B_j - g_j| \leq c_j \sec(\pi/p) \ , \ j = 1,\ldots,\ell \tag{16}$$

$$|z'_j - a_j| \leq d_j \sec(\pi/p) \ , \ j = 1,\ldots,n \tag{17}$$

Proof. Inequalities (15) - (17) are all established in the same way. For example, since we must have $|z'_j - a_j|_\Theta \leq d_j$, it follows from (12) that

$$|z'_j - a_j| \leq |z'_j - a_j|_\Theta \sec(\pi/p) \leq d_j \sec(\pi/p.)$$

The following sequence of inequalities establishes (14):

$$\begin{aligned}
\varepsilon' &\overset{\Delta}{=} \max |z'A_j - f_j|_\Theta \\
&\leq \max |z''A_j - f_j|_\Theta \\
&\leq \max |z''A_j - f_j| \overset{\Delta}{=} \varepsilon'' \\
&\leq \max |z'A_j - f_j| \\
&\leq \max |z'A_j - f_j|_\Theta \qquad \sec(\pi/p) \\
&= \varepsilon' \sec(\pi/p)
\end{aligned}$$

where the max in all cases is over $j = 1,\ldots,m$. This concludes the proof.

This theorem proves that the solution of the linearized problem is an approximate solution of the original nonlinear problem (1) - (4). Furthermore, it proves that the maximum relative error in this approximate solution can be made as small as desired by appropriate initial choice of p.

The original problem (1) - (4) has a mathematically straightforward solution if all the quantities are real-valued instead of complex; that is, the real-valued problem is an ordinary linear program in n+1 variables with $2(m+n+\ell)$ inequality constraints which can be solved in a finite number of steps using extensions of existing methods (see, e.g., [2], [3]). The complex-valued problem is much less simple. Eliminating complex arithmetic from the problem by substituting in the real and imaginary parts of all complex quantities yields (after squaring all the constraints) a mathematical programming problem in 2n+1 variables, with a linear objective function and $m+n+\ell$ quadratic constraints. No method is known which can solve this problem in a finiue number of steps. Since it is a convex programming problem and all functions involved have easily obtained derivatives of all orders, a great many different algorithms are potentially applicable for its approximate solution. However, the only reference [4] known to the author which explicitly studies the problem (1) - (4) uses a feasible directions method. At each step, a linear program is solved to determine the steepest feasible descent direction, a line search determines the step length, and special precautions are taken to prevent zig-zagging, or jamming. Also, a convergence proof is supplied. This method is not pursued further here.

The problem (1) - (4) can be viewed as a semi-infinite program (SIP), as is discussed briefly in section III. The SIP formulation of the problem consisting of the objective function (1) with only the constraints (2) has been studied in [5], [6], and [7], where it appears in the form of complex function approximation. The linearized absolute value (6) was first given in [6]. Theorem 1 and a less general form of Theorem 2 were first given in [7].

II. Solution of the linearized problem

A solution algorithm for the linearized problem for fixed p is discussed in this section. Special attention is directed to aspects of the problem which give rise to interesting special structure as a result of the underlying complex arithmetic of the problem. Several useful theoretical results are interspersed.

It is first established that the linearized problem (7) - (10) is a linear program, as stated in the Introduction. Denote the real and imaginary parts of any quantity x by x^R and x^I, respectively, whether x be a number, a row or column vector, or a matrix. By definition (6),

$$|zA_j - f_j|_\Theta = \max_{\theta\epsilon\Theta} [(zA_j - f_j)^R \cos\theta + (zA_j - f_j)^I \sin\theta],$$

so the m inequalities (8) evidently are equivalent to the system of mp inequalities

$$(zA_j - f_j)^R \cos\theta + (zA_j - f_j)^I \sin\theta \le \varepsilon \;,\; \theta \in \Theta \;,\; j = 1,\ldots,m \;. \qquad (18)$$

Since

$$(zA_j - f_j)^R = z^R A_j^R - z^I A_j^I - f_j^R \tag{19}$$

$$(zA_j - f_j)^I = z^R A_j^I + z^I A_j^R - f_j^I \quad ,$$

it is convenient to write (18) in the form

$$[z^R \; z^I \; \varepsilon] \begin{bmatrix} A^R \cos\theta + A^I \sin\theta \\ A^R \sin\theta - A^I \cos\theta \\ -1_m \end{bmatrix} \le [f^R \cos\theta + f^I \sin\theta], \; \theta \in \Theta, \tag{20}$$

where we have defined $1_m \in R^m$ to be the row vector whose components all equal one. Thus for each θ in (20) there corresponds m inequalities. The inequalities (9) and (10) may be treated similarly, so the linearized problem is a linear program in $2n+1$ variables and $m+n+\ell$ inequalities as claimed. The linear program can be written explicitly as follows:

Primal Problem.

$$\min_{[z^R \; z^I \; \varepsilon] \in R^{2n+1}} [z^R \; z^I \; \varepsilon] \begin{bmatrix} 0_n \\ 0_n \\ 1 \end{bmatrix} \tag{21}$$

subject to: $\epsilon \geq 0$ and, for each $\theta \in \Theta$,

$$[z^R \; z^I \; \epsilon] \begin{bmatrix} A^R \cos\theta + A^I \sin\theta & B^R \cos\theta + B^I \sin\theta & I \cos\theta \\ A^R \sin\theta - A^I \cos\theta & B^R \sin\theta - B^I \cos\theta & I \sin\theta \\ -1_m & 0_\ell & 0_n \end{bmatrix} \qquad (22)$$

$$\leq [f^R\cos\theta + f^I\sin\theta \quad c + g^R\cos\theta + g^I\sin\theta \quad d + a^R\cos\theta + a^I\sin\theta].$$

We employ the notation $I \in R^{n\times n}$ for the identity matrix and 0_k to denote either a zero row or a zero column of length $k \geq 1$. Whether it denotes a row or column will be clear from the context.

The primal problem is solved by solving its dual using the revised simplex method. The simplex (Lagrange) multipliers for an optimal basic solution of the dual problem provide a solution of the primal, assuming it to be feasible. The dual can be written in one of the standard linear programming formats by explicitly adding a slack variable, denoted Q, which arises naturally in this problem.

Dual Problem.

$$\min_{\substack{S \in R^{m\times p},\, T \in R^{\ell\times p} \\ W \in R^{n\times p},\, Q \in R}} \sum_{k=1}^{p} \left\{ \begin{array}{c} (f^R \cos\theta_k + f^I \sin\theta_k)\, S_k \\ + (c + g^R \cos\theta_k + g^I \sin\theta_k)\, T_k \\ + (d + a^R \cos\theta_k + a^I \sin\theta_k)\, W_k \end{array} \right\} \qquad (23)$$

subject to: $S \geq 0$, $T \geq 0$, $W \geq 0$, $Q \geq 0$, and

$$\sum_{k=1}^{p} \begin{bmatrix} A^R\cos\theta_k + A^I\sin\theta_k & B^R\cos\theta_k + B^I\sin\theta_k & I\cos\theta_k \\ A^R\sin\theta_k - A^I\cos\theta_k & B^R\sin\theta_k - B^I\cos\theta_k & I\sin\theta_k \\ 1_m & 0_\ell & 0_n \end{bmatrix} \begin{bmatrix} S_k \\ T_k \\ W_k \end{bmatrix} + \begin{bmatrix} 0_n \\ 0_n \\ 1 \end{bmatrix} Q = \begin{bmatrix} 0_n \\ 0_n \\ 1 \end{bmatrix} \quad (24)$$

Applying the asymmetric form of the standard duality relationship (see, e.g., [8, p. 69]), it is easy to verify that the dual of the dual problem (23) - (24) is equivalent to the primal problem (21) - (22). An alternative statement of the dual problem is given at the end of this section.

The slack variable Q plays a special role, as seen in the next result.

Theorem 3. Let the matrices $S' \geq 0$, $W' \geq 0$, $T' \geq 0$, and the real number $Q' \geq 0$ denote any optimal basic feasible solution of the dual problem (23) - (24). If $Q' > 0$, then the optimal value of the objective function in the primal problem is zero.

Proof. The proof uses the asymmetric form of the Complementary Slackness Theorem (see, e.g., [8, p. 77]). Let $[z'^R \; z'^I \; \epsilon'] \in R^{2n+1}$ denote the simplex multipliers corresponding to the optimal basic solution S', W', T', Q'. By the Complementary Slackness Theorem, $Q' > 0$ implies that

$$[z'^R\ z'^I\ \varepsilon'] \begin{bmatrix} 0_n \\ 0_n \\ 1 \end{bmatrix} = \varepsilon' = 0.$$

This completes the proof.

Except for the slack variable Q, every basic variable of the dual problem is uniquely identified by specifying the matrix to which it belongs together with its location (row and column number) in this matrix. The matrix names S, T, and W clearly correspond to the inequality systems (8), (9), and (10), respectively, of the linearized problem. The row number of a basic variable identifies the particular constraint which gave rise to it. For example, all the dual variables in row q of matrix T would be eliminated from the dual problem if the q-th inequality in (9) were deleted from the linearized problem. Similarly, the column.number of a basic variable identifies the angle in the set Θ to which it may be said to correspond. Thus, adjoining a new angle to the set Θ causes one new column to be augmented to each of the matrices S, T, and W.

The revised simplex algorithm, as applied to the dual problem, is defined in general terms as follows:

<u>Step 1</u>. Determine an initial basic feasible solution of the dual problem.

<u>Step 2</u>. Compute the simplex multipliers corresponding to the current basic feasible solution.

<u>Step 3</u>. Determine the incoming variable by selecting the variable having the most negative reduced cost coefficient; terminate if all reduced cost coefficients are nonnegative — the primal problem is solved by the current simplex multipliers.

Step 4. Compute the column of the incoming variable in terms of the current basis.

Step 5. Determine the outgoing basic variable by a ratio test; terminate if the dual objective function is unbounded below -- the primal problem is infeasible.

Step 6. Update the basis inverse and current basic feasible solution by pivoting, and return to Step 2.

The special structure of the dual problem has its strongest influence on Steps 1, 3, and 4. These effects are outlined next. Discussion of other important aspects of the algorithm are postponed to section IV.

The dual problem is already in canonical form for initiating the second phase of the simplex algorithm. In other words, Step 1 is trivial. To see this, we need only show that a (2n+1) × (2n+1) identity matrix can be assembled from the columns of the coefficient matrix of (24). One readily available column is the column corresponding to the slack variable Q. The remaining 2n columns correspond to dual variables which are the components of two particular W columns. Recalling the construction of the set Θ, we have $\theta_1 = 0$ so that $\cos\theta_1 = 1$ and $\sin\theta_1 = 0$. Hence one of these W columns can be taken to be W_1. Similarly, the other is $W_{1+p/4}$ since $\theta_{1+p/4} = \pi/2$. Hence the initial basic feasible solution is

$$\begin{bmatrix} W_1 \\ W_{1+p/4} \\ Q \end{bmatrix} = \begin{bmatrix} 0_n \\ 0_n \\ 1 \end{bmatrix} . \qquad (25)$$

The simplex multipliers corresponding to (25) can be derived in the standard manner; however, it is more convenient to derive them in a special way later in this section.

The remark was made in the Introduction that the constraints (10) would be present implicitly even when not explicitly included in the linearized problem. If the constraints (10) are deleted, then all of the W variables must be deleted from the dual problem. An easily recognized initial basic feasible solution would not then be available, so it would be necessary to incorporate $2n$ artificial variables into the dual problem. These artificial variables, together with the slack variable Q, would constitute an initial basic feasible solution. It is easy to see that this solution is, in fact, indistinguishable from (25); hence, the artificial variables are actually the two columns W_1 and $W_{1+p/4}$ in disguise. The only difference is that the artificial variables would, of course, have zero cost coefficients.

The initial basic feasible solution (25) is highly degenerate because all but one of the constant terms in (24) are zero. As discussed in [9], it is in problems of this general kind that cycling in the simplex algorithm has been occasionally observed in practice. Such cycling was observed in an example given in this report. However, a trivial modification of the tie-breaking rule in the ratio test for the outgoing basic variable, together with "preferential treatment" of certain incoming variables, seems to completely avoid the problem. Further discussion of cycling in the dual problem is postponed to section IV.

The cost coefficients and the columns of any dual variable can be found by inspection of (23) - (24). It is better, however, to think of them in the complex arithmetic format given in Table I, which is

dual variable	cost coefficient	column, in R^{2n+1}
S_{jk}	$\left(f_j e^{-i\theta_k}\right)^R$	$\begin{bmatrix} \left(A_j e^{-i\theta_k}\right)^R \\ -\left(A_j e^{-i\theta_k}\right)^I \\ 1 \end{bmatrix}$
T_{jk}	$\left(c_j + g_j e^{-i\theta_k}\right)^R$	$\begin{bmatrix} \left(B_j e^{-i\theta_k}\right)^R \\ -\left(B_j e^{-i\theta_k}\right)^I \\ 0 \end{bmatrix}$
W_{jk}	$\left(d_j + a_j e^{-i\theta_k}\right)^R$	$\begin{bmatrix} \left(I_j e^{-i\theta_k}\right)^R \\ -\left(I_j e^{-i\theta_k}\right)^I \\ 0 \end{bmatrix}$

Table 1. Dual variable cost coefficients and columns in complex arithmetic format.

readily verified. Utilizing this format makes a very significant reduction in computer storage easy to understand. Instead of storing the $(m+n+\ell)p$ columns of all the dual variables, the column of a dual variable is constructed only if it is needed. As indicated in Table 1, this requires storing only the matrices A and B and the cosines and sines of each of the p angles in Θ. Assuming the necessary 2p cosines and sines are computed once and for all at the outset, only n complex multiplications are required to construct the column of any specified dual variable. This is a very small price to pay for reducing the storage from $(2n+1)(m+n+\ell)p$ words to only $2n(m+n+\ell)+2p$ words. Furthermore, the columns of the dual variables W_{jk} are constructed from the identity matrix I, which need not be explicitly stored. Hence the total storage necessary for constructing the column of any dual variable is merely $2n(m+\ell)+2p$ words.

A very efficient method of computing the smallest reduced cost coefficient in Step 3 of the revised simplex algorithm is now discussed. This method is particularly interesting because none of the columns of the dual variables are explicitly needed. The only data required are the original complex matrices A and B and the sines and cosines of the angles in Θ . Let λ be any real row vector of simplex multipliers for the dual problem; thus, λ is of length $2n+1$. The vector λ defines a complex row vector $z \in C^n$ and a real number $\varepsilon \in R$ by the identification

$$\lambda \triangleq [z^R \quad z^I \quad -\varepsilon] \in R^{2n+1} \text{ .} \qquad (26)$$

This definition is reasonable considering the statement (21) - (22) of the primal problem. The reduced cost of the dual variable S_{jk} is the cost coefficient of S_{jk} minus the row λ times the column of S_{jk}. Using (26) and Table 1 shows that

$$C_S^{jk} \triangleq \left(f_j e^{-i\theta_k}\right)^R - \left[\, z^R \; z^I \; -\varepsilon \,\right] \begin{bmatrix} \left(A_j e^{-i\theta_k}\right)^R \\ -\left(A_j e^{-i\theta_k}\right)^I \\ 1 \end{bmatrix}$$

$$= \varepsilon - \left[\left(zA_j - f_j\right) e^{-i\theta_k} \right]^R , \qquad (27)$$

so the minimum reduced cost coefficient of all p variables in the j-th row of S must be

$$C_S^j \triangleq \min_{1\leq k\leq p} C_S^{jk} = \varepsilon - \left|zA_j - f_j\right|_\Theta , \quad j = 1,2,\ldots,m. \qquad (28)$$

The smallest reduced cost coefficient of all the dual variables of the matrix S is then

$$C_S \triangleq \min_{1\leq j\leq m} C_S^j = \varepsilon - \max_{1\leq j\leq m} \left|zA_j - f_j\right|_\Theta . \qquad (29)$$

Similarly, the minimum reduced cost coefficients over all the dual variables of T and W are

$$C_T \triangleq \min_{1\leq j\leq \ell} \left(c_j - \left|zB_j - g_j\right|_\Theta\right) \qquad (30)$$

and

$$C_W \overset{\Delta}{=} \min_{1 \leq j \leq n} \left\{d_j - |z_j - a_j|_\Theta\right\} \quad , \tag{31}$$

respectively. The smallest reduced cost of all the variables of the dual problem is

$$C_{SWT} \overset{\Delta}{=} \min \{C_S, C_W, C_T\}. \tag{32}$$

The actual value of the minimum reduced cost C_{SWT} is not particularly important. What is important is knowing for which dual variable the minimum is attained. This requires knowing which of the three quantities C_S, C_W, or C_T is smallest as well as knowing for which index j the minimum values of (29) - (31) are attained. These two pieces of data tell the row number and the correct matrix name of the incoming dual variable. To determine the column number, we must also know which angle $\theta_k \in \Theta$ gives the largest projection (i.e., the linearized absolute value (6)) at the minimal index j. Since the angle θ_k may not be unique because of possible ties in (6), a consistent tie-breaking rule must be defined. This rule, which we will call the minimal clockwise index (MCI) rule, is important because it determines unambiguously the name of the incoming dual variable in the simplex method.

Let $u \in C$, and let u_Θ be the set of those angles $\theta \in \Theta$ for which the maximum in (6) is attained. There are three cases. First, if u_Θ has precisely one element, the MCI of u is defined to be the index of that element. Second, if u_Θ has precisely two elements, say θ_k and θ_j, and neither k or j equals p, then the MCI of u is defined to be

$\min \{k,j\}$; on the other hand, if either $k = p$ or $j = p$, then the MCI of u is taken to be p. The reason for the exception is clear from the geometry of the problem. Third, if u_Θ has more than two elements, then it must be that $u = 0$ and $u_\Theta = \Theta$, so the MCI of u is defined to be 1.

Thinking of the simplex multipliers as a complex vector of length n and a real number, as in (26), is "natural" in this problem. The complex number $zA_j - f_j$, for example, can be calculated entirely in complex arithmetic. This makes any computer program implementing the algorithm easier to write and also probably makes it execute more efficiently (when coded intrinsically in COMPLEX mode). Furthermore, the idea emerges that the computation of the linearized absolute value $|zA_j - f_j|_\Theta$ and its MCI should be treated as a separate optimization problem having its own special features. This subproblem is discussed next.

The computation of the linearized absolute value and corresponding MCI must be undertaken for $m+n+\ell$ complex numbers during each iteration of the simplex algorithm in Step 3. A brute force approach using the definition (6) might require as many as $2p$ real multiplications for each complex number. Such an approach is quite inefficient and does not exploit the special form of the set Θ. For $p = 4$, it is clear that comparison tests alone suffice to solve this subproblem. For $p \geq 8$, we claim that comparison tests and at most $2 \log_2 p - 5$ real multiplications are sufficient. To see this, first determine the octant of the complex plane in which the given number lies, and then determine whether it lies above or below the 45° line bisecting the octant. This can be done using comparison tests only. Now that the "half-octant" in which the number lies is known, its projections onto the bounding rays of this half-octant can be computed in this special case using only one multiplication. If $p = 8$, a final comparison test ends the problem. If $p \geq 16$, then the larger of the two projections reveals the "quarter-octant" in which the number must lie. The projection onto one of the bounding rays of this quarter-octant is already known, so it is only necessary to compute the projection onto the other bounding ray. This requires 2 real multiplications. If $p = 16$, a final comparison test ends the problem. If $p \geq 32$, we continue as before. Counting the total possible number of steps proves our claim. This bisection method works because we have required the set Θ to contain only the points (11). Other choices of Θ would require different methods.

The number of real multiplications required to complete Step 3 of the revised simplex algorithm using the methods discussed above is significantly less than that required in the usual approach. Given the simplex multipliers, the straightforward method requires the computation of $(m+n+\ell)p - (2n+1)$ real inner products of length 2n. Taking account of the simple form of the W columns gives a total of approximately

$$(2p-4)[n(m+\ell)+1] + 4n(m+\ell-n-1/2) \tag{33}$$

real multiplications. The special methods discussed above require $m+n+\ell$ complex inner products of length n followed by $2 \log_2 p - 5$ real multiplications for each inner product. Counting one complex multiplication as four real multiplications and considering the special form of the W columns gives a total of

$$4n(m+\ell) + (2 \log_2 p - 5)(m+n+\ell) \tag{34}$$

real multiplications. Clearly the special methods are substantially better than the straightforward methods when $p \geq 8$ and $m > n$. Even for $p = 4$, the special methods are superior because the factor $2 \log_2 p - 5$ in (34) must be replaced by zero. In the derivation of both (33) and (34) it was assumed that the last row of (24) in the dual problem was specially treated to avoid multiplications by 1 and 0. Also, (33) is valid for all $p \geq 3$, while (34) is valid only for $p = 8,16,32,\ldots$.

We earlier postponed the derivation of simplex multipliers $\lambda^{(0)} \in R^{2n+1}$ corresponding to the initial basic feasible solution (25). They are now derived in the complex format (26). Multiplying the initial basis inverse on the left by the row vector containing the cost coefficients of the initial basic variables gives the row vector $\lambda^{(0)}$. The initial basis inverse is the identity matrix, the cost coefficients of the basic W variables are given in Table 1, and the cost coefficient of the slack variable Q is 0. Since $\theta_1 = 0$ and $\theta_{1+p/4} = \pi/2$, we have

$$\lambda^{(0)} = (d+(a)^R,\ d+(-ia)^R,\ 0) = (d+a^R,\ d+a^I,\ 0) \in R^{2n+1} \ .$$

The definition (26) thus gives

$$z^{(0)} \triangleq a + d\ e^{i\pi/4} \sqrt{2} \in C^n \quad , \quad \epsilon^{(0)} \triangleq 0 \ . \tag{35}$$

From the proof of Theorem 3 it can be seen that $\varepsilon = 0$ for as long as the slack variable remains in the basis and is positive.

The only sparse matrices in the dual problem are the matrices S, W, and T. Any basic feasible solution of the dual consists of $2n+1$ variables, all of which must be non-negative. Since the remainder of the $(m+n+\ell)p + 1$ dual variables must be zero, the matrices S, W, and T must be sparse. Furthermore, no row of S, W, and T need contain more than two positive entries as the next theorem shows.

Theorem 4. No basic feasible solution of the dual problem (23) - (24) can have more than two basic variables in any one row of W or T. If a basic feasible solution of the dual problem has corresponding simplex multipliers (26) with $\varepsilon > 0$, then it cannot have more than two basic variables in any one row of S.

Proof. We prove the first statement for the matrix T, since the proof for the W matrix is just a special case. Consider the j-th row of T. Suppose a basic feasible solution exists which has the three basic variables T_{jr}, T_{js}, and T_{jt} with r, s, and t being distinct. Then the corresponding simplex multipliers $z \in C^n$ and $\varepsilon \in R$ must result in zero reduced costs for all three variables. A result analogous to (27) was used to prove (30); using that analogous result here too gives

$$C_T^{jq} = 0 = c_j - \left[(zB_j - g_j)\, e^{-i\theta_q} \right]^R , \quad q = r,s,t . \qquad (36)$$

Thus the single complex number $zB_j - g_j$ must have the same projection, namely c_j, in each of three distinct directions. This is impossible unless $zB_j - g_j = c_j = 0$, in contradiction to our initial assumption that $c_j > 0$. This establishes the first statement of the theorem. The second theorem statement is proved in the same way, by using (27) itself. This concludes the proof.

The following theorem relates knowledge of an optimal basis of the dual to "observable" quantities in the primal problem. The results of the theorem depend on the names, but not the actual numerical values, of the optimal basic variables. In addition it seems to indicate that the upper bound (14) in Theorem 2 will often be attained in practice.

Theorem 5. Let $\varepsilon' \in R$ and $z' \in C^n$ denote the simplex multipliers in the form (26) of a given optimal basis for the dual problem (23) - (24), and suppose that $\varepsilon' > 0$. If the j-th row of one of the matrices S, W, or T contains two optimal basic variables in columns r and t with $p \geq r > t \geq 1$, then either $r - t = 1$ or $r - t = p-1$. If $r - t = 1$, then

$$z'A_j - f_j = \varepsilon' \sec(\pi/p) \exp[i(2t - 1)\pi/p] \tag{37}$$

$$z'B_j - g_j = c_j \sec(\pi/p) \exp[i(2t - 1)\pi/p] \tag{38}$$

$$z_j' - a_j = d_j \sec(\pi/p) \exp[i(2t - 1)\pi/p], \tag{39}$$

according to whether the j-th row is a row of S, T, or W, respectively. Replacing t with p in (37) - (39) gives the equations corresponding to the alternative case $r - t = p - 1$.

Proof. We treat only the S matrix case since the other two cases are very similar. Thus, the two basic variables involved are S_{jr} and S_{jt}, and we assume that $p \geq r > t \geq 1$. The reduced costs C_S^{jr} and C_S^{jt} must be 0, so from (27)

$$\begin{aligned} \varepsilon' &= \left[(z'A_j - f_j)e^{-i\theta_r} \right]^R \\ \varepsilon' &= \left[(z'A_j - f_j)e^{-i\theta_t} \right]^R . \end{aligned} \tag{40}$$

Any complex number having identical projections in two directions is uniquely defined in both magnitude and phase. If θ_r differs from θ_t by π radians, the system (40) implies that $\varepsilon' = 0$, contrary to assumption. Thus θ_r and θ_t differ by more or less than π, and (40) implies that

$$\left| z'A_j - f_j \right| = \varepsilon' \sec(\psi/2) > 0 ,$$

where $\psi = \min \{\theta_r - \theta_t , 2\pi - \theta_r + \theta_t\}$. By Theorem 2, ψ must equal π/p, so that either $\theta_r - \theta_t = \pi/p$ or $\theta_r - \theta_t = \pi(2p - 1)/p$. From (11) we have either $r - t = 1$ or $r - t = p - 1$. For $r - t = 1$, solving the system (40) for the phase of $z'A_j - f_j$ gives (37). The case $r - t = p - 1$ is handled in the same way. This completes the proof.

Theorem 5 is useful in practice as well. Computed optimal basic solutions can be inspected very easily to verify that the optimal basic variables occurring in the same row are in fact "paired" in the manner described. If they are not, then one must conclude that premature termination occurred in the simplex algorithm, or else that numerical round-off errors have adversely affected the computed solution.

Theorem 6. Let ε' and z' be as in Theorem 5. If the j-th row of one of the matrices S, W, or T contains an optimal basis variable in column r, $1 \leq r \leq p$, then

$$\begin{aligned} &\varepsilon' \leq |z'A_j - f_j| \leq \varepsilon' \sec \pi/p \\ &\theta_r - \pi/p \leq \arg (z'A_j - f_j) \leq \theta_r + \pi/p, \end{aligned} \qquad (41)$$

or

$$\begin{aligned} &c_j \leq |z'B_j - g_j| \leq c_j \sec \pi/p \\ &\theta_r - \pi/p \leq \arg (z'B_j - g_j) \leq \theta_r + \pi/p, \end{aligned} \qquad (42)$$

or

$$\begin{aligned} &d_j \leq |z_j - a_j| \leq d_j \sec \pi/p \\ &\theta_r - \pi/p \leq \arg (z_j - a_j) \leq \theta_r + \pi/p, \end{aligned} \qquad (43)$$

according to whether the j-th row is a row of S, W, or T, respectively.

Proof. The proof is closely related to the method of proof of Theorem 5, and it is not presented. We remark that if Theorem 6 were proved first, then Theorem 5 would be an immediate consequence of it.

This section is concluded with an alternative statement of the dual problem (23) - (24) using complex arithmetic notation. We do not use it elsewhere in this report, but this form is interesting in its own right and also provides a more concise statement of the dual which may be useful.

Dual Problem: complex format

$$\min_{\substack{S,T\\W,Q}} \quad \left[(fS + gT + aW)e^{-i\Theta}\right]^R + \sum_{j=1}^{p}(CT_j + DW_j)$$

subject to: $S \geq 0$, $T \geq 0$, $W \geq 0$, $Q \geq 0$, and

$$(AS + BT + W)e^{-i\Theta} = 0 \in C^n$$

$$Q + \sum_{j=1}^{m} \sum_{k=1}^{p} S_{jk} = 1 \in R.$$

We have used $e^{-i\Theta}$ to denote a complex column vector of length n whose k-th component is $e^{-i\theta_k} \in C$; all other notation is unchanged from (23) - (24). Verifying that this complex format is correct is straightforward.

III. Solution of the linearized problem for large p.

The relative accuracy between the absolute value and the linearized absolute value is given by (12). Although in some applications a choice of $p = 16$ may provide satisfactory accuracy, in other applications $p = 16$ may be much too small. As stated in the Introduction, $p = 1024$ gives 5 significant digits of relative accuracy, so applications requiring greater accuracy will therefore need $p > 1024$. There is, however, a practical limit to how large p may be taken in many problems. For example, a problem will become numerically unstable for sufficiently large values of p if its optimal solution has, for every p, two basic dual variables in at least one row of S W, or T. From Theorem 5, these problems have two "consecutive" projections in an optimal basis and, since the inequalities defining these projections become progressively less distinguishable numerically as p increases, the basis matrix must become progressively more ill-conditioned. Only those problems which never, for any p, have more than one optimal basic variable in any one row of S, W, or T will escape numerical ill-conditioning from this cause; however, problems of this kind seem to be uncommon.

It is useful, nonetheless, to be able to solve the linearized problem for as large a value of p as is numerically practical. One reason is that a solution of the linearized problem furnishes a starting point for other methods which potentially provide greater accuracy. For instance, the problem (1) - (4) can be rewritten as a semi-infinite program, or SIP, and an interesting algorithm [10], [11] for solving a class of general SIP's can be utilized to solve it. This method sets up an appropriate nonlinear system of algebraic equations which are then solved using the Newton-Raphson method (or other workable iterative method). A feasible solution of this nonlinear system gives a solution of the SIP. The starting point of the Newton-Raphson iteration can be taken to be the solution of the linearized problem (7) - (10). Conceivably a very good starting point may be needed to ensure that the Newton-Raphson iteration converges to a feasible point. In this event, the linearized problem can be made more accurate merely be increasing p. The subproblem (1) - (2) is treated as an SIP in [5], alghough no mention is made there of special structure in the linearized problem or of results such as those in any of the theorems above.

The method we suggest for solving the linearized problem for large values of p begins by solving the smallest dual problem. That is, the dual problem (23) - (24) with $p = 4$ is solved by the simplex method starting at the basic feasible solution (25). Next, the $p = 8$ dual problem is solved using the optimal basis for the $p = 4$ dual to start the simplex algorithm. The $p = 16$ dual is then solved starting at the optimal basis for the $p = 8$ dual. Continuing in this fashion, large values of p are quickly reached. The algorithm is always well-defined because basic feasible solutions of the dual for a given p are also basic feasible solutions of the dual for all larger values of p. This special feature of the dual holds because the Θ sets are nested for the allowed values $p = 4, 8, 16, 32, \ldots$.

By doubling p at each stage beginning with $p = 4$, this algorithm permits the simplex iterations to avoid bases associated with numerical instability from the linearization process until p becomes large. In other words, potentially troublesome bases are not encountered until a significant number of simplex iterations have already transpired. Therefore greater numerical accuracy might be anticipated. The algorithm, of course, cannot avoid difficulties caused by ill-conditioning in the complex equations themselves.

This method is practical for large values of p only because the methods presented in section II require computer storage that is almost independent of p. As pointed out there, the dual problem requires only $2n(m+\ell) + 2p$ storage locations. With an explicit inverse formulation of the simplex method, an additional $(2n+1)^2$ words are required. Thus the only storage dependent on p are the 2p locations for storing the cosines and sines of the projections. (Actually, only p/8 storage locations can be made to suffice for this purpose.)

One advantage of this algorithm is that the optimal basis for each intermediate value of p can be easily inspected using Theorems 5 and 6 to determine to what extent numerical round-off errors are present in the solution. If sufficient error is present, the algorithm can be terminated early, or alternatively, the basis can be reinverted before continuing to the next value of p. Either way, pointless simplex iterations can be avoided.

The only drawback the algorithm potentially has is that more simplex iterations might be required to reach the final optimal dual basis by forcing it to proceed via smaller values of p than by solving the full dual problem all at once. This difficulty does not seem to be significant in practice. Should it ever become a problem, however, it could be at least partially overcome by skipping more rapidly through the available values of p. For instance, rather than doubling p at each stage, one could quadruple it instead. It is also possible to begin the algorithm with a larger initial value of p; that is, $p > 4$.

An incidental advantage of the algorithm is that the primal solutions for all the intermediate values of p can be stored. This raises the possibility of solving the original problem (1) - (4) by extrapo'ation. The original problem corresponds to the case $p = \infty$, so it is not unreasonable to think that better solutions could be obtained by extrapolating the solutions corresponding to the finite cases $p = 4, 8, 16, 32, \ldots$. The optimal vectors $\dot{z}$ of the linearized primal problems converge linearly with increasing p, while the optimal values ε converge quadratically with increasing p. Therefore Richardson extrapolation (see, for instance [12] or [13]) might be suitable for extrapolating primal solutions. Extrapolating the $p = \infty$ dual solution [14] from the finite p dual solutions requires special care. The optimal finite dual basis names must be normalized before extrapolation, i.e., the column numbers of the basic variables must be multiplied by $2\pi/p$ to map them to the interval $(0,2\pi)$, and the possible discontinuity at 2π must be taken into account during extrapolation. Also, those optimal finite p dual basic variables which occur in "pairs", as described in Theorem 5 above, must coalesce into one optimal $p = \infty$ dual basic variable.

IV. Details of the revised simplex algorithm

It is not desirable to use a computer code which treats the complex matrices and vectors of the primal problem by separating them into their real and imaginary parts. Besides the inconvenience, such a code would be very inefficient in practice because the reduced cost coefficient calculations (29) - (30) would cause thrashing on virtual memory systems. (See [15] for an example involving Hermitian eigenproblem solution.) Given, then, that the solution vector z of the primal problem is best stored as a complex vector, it becomes clear that the simplex multipliers (26) should be re-ordered to reflect the storage of z. Consequently the rows (24) of the dual problem should also be re-ordered. The computer code therefore visualizes the dual problem rows in the following order:

$$\{1,\ n+1,\ 2,\ n+2,\ \ldots,\ n-1,\ 2n-1,\ n,\ 2n,\ 2n+1\} \tag{44}$$

where these numbers denote the row numbers in the original system (24). The re-ordered system turns out to be much easier to work with in FORTRAN code than the original system. With the rows of the dual problem ordered as in (44), the reduced cost calculations can be coded in FORTRAN just as they are written in (29) - (31), provided the initial data of the problem are typed COMPLEX.

The most negative reduced cost (32) determines the entering basic variable in the simplex algorithm. If no reduced cost is negative, the simplex algorithm terminates because the current basis is optimal. If ties develop for the most negative reduced cost, they are broken by choosing the variable with the least lexicographical index. Since every dual variable has three names, the "index" of a dual variable is a triplet of positive integers:

$$i/j/k$$

where i = 1, 2, or 3 according to whether it is an S, W, or T variable

j = constraint number; from (8) - (10)

k = projection number of the angle in the set Θ; $1 \leq k \leq p$.

Note that the middle index j has different ranges of possible values depending on the value of i. These triplets are ordered lexicographically, so the least index is well-defined.

There is one exception to the least index rule in case of ties for the entering variable. The exception arises because the highly degenerate initial starting point (25) can cause cycling in the simplex algorithm. As long as the slack variable Q remains in the basis, the only entering variables permitted are S variables with negative reduced costs. If S

variables with negative reduced costs do not exist, then the entering variable is permitted to be a W or a T variable and the ties are resolved by the least index rule as described above. Thus, S variables are given priority for entering the basis only for as long as the slack Q is in the basis. Once Q is removed from the basis it is never allowed to re-enter, and the exception to the least index tie breaking rule becomes moot.

The outgoing basic variable is determined by the usual ratio test, with ties resolved by taking the variable having the maximum magnitude pivot with the smallest index. That is, the variable having the least ratio in the ratio test must leave the basis. If the least ratio is attained by more than one variable, the variable having the largest magnitude pivot is chosen. If more than one of these variables have the same magnitude pivot, then the variable with least index is selected.

Because of degeneracy and cycling, there is one exception to this tie breaking rule for the exiting variable. So long as the slack Q remains in the basis, only W variables are permitted to exit. This rule makes sense only when a W variable is involved in the tie; if no such W variable exists, the exception is not invoked. If more than one W variable is involved in the tie, then the one having the largest magnitude pivot with the least index is selected to exit. Just as for the entering variable, this exception becomes moot once the slack Q leaves the basis.

With these modifications to the usual tie breaking rules for entering and exiting variables, no cycling in the simplex algorithm has yet been observed. Hcwever, if these modifications are deleted, cycling may very well occur. Example 3 of section V below cycled (with a cycle of length 19) without these modifications. Whether or not cycling in this particular example still happens when exact arithmetic is used is not known to the author. It is possible that the observed cycling is merely an artifact of finite precision arithmetic.

In practical implementations of the simplex algorithm, a zero tolerance is necessary when testing for the most negative reduced cost and for possible divisors in the ratio test. This number is somewhat arbitrary, but it is important that it not be too small and that it somehow be dependent on the scale of the data of the problem. The number used here is the product of the unit round-off error of the host computer and the sum of the absolute values of the incoming column (i.e., its L_1 norm). This number is rather conservative but seems satisfactory in the examples run to date. It is used for both the reduced cost and the pivot tolerance tests.

Besides the usual expected termination criteria in the simplex algorithm, the pricing method implicit in (29) - (31) yields an interesting and perhaps novel way to terminate the algorithm. The pricing method computes the most negative reduced cost by examination of <u>all</u> the reduced costs, not merely the reduced costs of the nonbasic variables. Hence it can happen that the entering and the exiting variable are identical because of numerical round-off errors. In practice this event seems to signal that no further improvement in the solution is numerically possible. Solutions returned by terminating the algorithm whenever this "self-cycling" occurs appear to be very satisfactory. On the other hand, failure to terminate in this situation leads to wasted simplex iterations. Eventually enough accumulated numerical round-off error develops to end self-cycling, but by then the solution is usually very poor.

A FORTRAN code was developed to implement and test the methods described for solving the dual problem. It holds an explicit basis inverse and performs the usual pivoting to update the inverse in each simplex iteration. This procedure is known to be numerically unstable, but easily programmed. To forestall numerical difficulties the inverse is held in double precision, even though a double precision inverse is not in general a satisfactory substitute for a numerically stable technique. Updating the LU or QR factorizations of the basis are preferable. Nonetheless the explicit inverse code gave good performance on the small problems tested in this report. Reinversion of the basis is also a desirable feature of a general purpose code, but it was not made part of the initial FORTRAN code.

V. Examples

Example 1 is taken directly from [4, p. 249]. It can be put in the form (1) - (3) by letting $n = 2$, $m = 5$, $\ell = 2$, and defining the matrices

$$A = \begin{bmatrix} 1 & 1 & 1 & .5 & 2 \\ -2 & 0 & 3 & -1 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} 2 & 2 \\ 2 & -4 \end{bmatrix} , \quad g = \begin{bmatrix} 0 \\ 0 \end{bmatrix} , \; c = \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} , \tag{45}$$

$$f = \begin{bmatrix} -1 + i \\ -1 + i \\ .5i \\ 0 \\ -1 + i \end{bmatrix} .$$

In this example all the data happen to be real valued, except the vector f. Since constraints of type (4) are missing from this example and since their linearizations provide the initial dual basis, we must insert them here. We take

$$a = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad d = \begin{bmatrix} 10 \\ 10 \end{bmatrix} \tag{46}$$

so that they will be inactive at the solution. The exact solution of this problem is $z_1 = (-1 + i)/2$, $z_2 = 0$, and $\varepsilon = \sqrt{2}/2$

Table 2 gives the solutions of the linearized primal problem for $p = 4, 8, 16,\ldots,2048$. It is apparent that the optimal value of ε for $p = 8$ is also the optimal value of ε for all $p \geq 16$. For $p \geq 8$, then, the accuracy of the primal solutions depends solely upon the linearization errors since the optimal ε does not change. Furthermore, the optimal value of the original nonlinear problem must surely equal the optimal ε for $p = 8$.

Table 3 gives the optimal basic solutions of the dual problems for $p = 4, 8,\ldots,2048$. The active constraints are the same for all $p \geq 8$, except for their θ names. Hence the active constraints at the optimum of the original nonlinear problem (1) - (4) have surely been identified. The fourth and fifth basic variables are "paired" in an obvious way; this behavior is explained by Theorem 5.

All the optimal solutions presented in Table 3 are degenerate, or very nearly so. It is particularly interesting that the "degenerate parts" of the optimal solutions almost double in size as p is doubled, especially when $p \geq 64$. Assuming this trend continues, the optimal solution will eventually look as if it is not degenerate. We suspect, but have not attempted to prove, that this trend is purely an artifact of the numerical ill-conditioning inherent in the linearization process, as was discussed in section III above.

Assuming linear convergence for increasing p, one step of the Richardson extrapolation process is easily applied to the last two z vectors in Table 2. Multiplying the $p = 2048$ vector by two and subtracting the $p = 1024$ vector gives

$$\begin{bmatrix} z_1^R \\ z_1^I \\ z_2^R \\ z_2^I \end{bmatrix} = \begin{bmatrix} -.500000 \\ +.500\mathrm{C}00 \\ +.001370 \times 10^{-11} \\ -.000481 \times 10^{-11} \end{bmatrix} .$$

Evidently one step of this extrapolation procedure has nearly doubled the number of correct significant digits in the primal solution.

p	4	8	16	32	64	128	256	512	1024	2048
z_1^R	-.588760	-.292893	-.400544	-.450754	-.475437	-.487726	-.493864	-.496932	-.498466	-.499233
z_1^I	.588760	.707107	.599456	.549246	.524563	.512274	.506136	.503068	.501534	.500767
z_2^R	$.591734 \times 10^{-1}$	$-.819967 \times 10^{-9}$	$-.278432 \times 10^{-9}$	$-.116877 \times 10^{-9}$	$-.534144 \times 10^{-10}$	$-.254916 \times 10^{-10}$	$-.124453 \times 10^{-10}$	$-.614763 \times 10^{-11}$	$-.305566 \times 10^{-11}$	$-.152098 \times 10^{-11}$
z_2^I	$-.591734 \times 10^{-1}$	$-.819967 \times 10^{-9}$	$-.115330 \times 10^{-9}$	$-.780945 \times 10^{-10}$	$-.438360 \times 10^{-10}$	$-.231043 \times 10^{-10}$	$-.118489 \times 10^{-10}$	$-.599856 \times 10^{-11}$	$-.301843 \times 10^{-11}$	$-.151162 \times 10^{-11}$
ε	.4112399	.7071068	.7071068	.7071068	.7071068	.7071068	.7071068	.7071068	.7071068	.7071068
Total Iterations	10	14	17	20	23	26	29	32	35	38

Table 2. Solutions of the linearized primal problem of Example 1.

p	4	8	16	32	64	128	256	512	1024	2048
basis names	1/2/1 1/2/4 1/3/3 3/2/2 3/2/3	1/1/8 1/2/1 3/1/4 3/2/3 3/2/4	1/1/15 1/2/16 3/1/6 3/2/6 3/2/7	1/1/29 1/2/30 3/1/12 3/2/12 3/2/13	1/1/57 1/2/58 3/1/24 3/2/24 3/2/25	1/1/113 1/2/114 3/1/48 3/2/48 3/2/49	1/1/225 1/2/226 3/1/96 3/2/96 3/2/97	1/1/449 1/2/450 3/1/192 3/2/192 3/2/193	1/1/897 1/2/898 3/1/384 3/2/384 3/2/385	1/1/1793 1/2/1794 3/1/768 3/2/768 3/2/769
basis values	.714286	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	.000000	$.503575 \times 10^{-19}$	$.657953 \times 10^{-19}$	$.109414 \times 10^{-18}$	$.199533 \times 10^{-18}$	$.380673 \times 10^{-18}$	$.743285 \times 10^{-18}$	$.146865 \times 10^{-17}$	$.291943 \times 10^{-17}$	$.582104 \times 10^{-17}$
	.285714	$-.498720 \times 10^{-37}$	$.908443 \times 10^{-19}$	$.243693 \times 10^{-19}$	$.545841 \times 10^{-19}$	$.115007 \times 10^{-18}$	$.235889 \times 10^{-18}$	$.477680 \times 10^{-18}$	$.961275 \times 10^{-18}$	$.192848 \times 10^{-17}$
	.000000	$.114997 \times 10^{-18}$	$.217029 \times 10^{-18}$	$.428992 \times 10^{-18}$	$.856894 \times 10^{-18}$	$.171471 \times 10^{-17}$	$.343135 \times 10^{-17}$	$.686516 \times 10^{-17}$	$.137330 \times 10^{-16}$	$.274689 \times 10^{-16}$
	.214286	.500000	.500000	.500000	.500000	.500000	.500000	.500000	.500000	.500000

Table 3. Optimal solutions of the dual problem of Example 1.

It is possible to make this problem infeasible by adjoining only one constraint of type (3). Instead of (45) we take

$$B = \begin{bmatrix} 2 & 2 & 1 \\ 2 & -4 & 1 \end{bmatrix}, \quad g = \begin{bmatrix} 0 \\ 0 \\ 7-4i \end{bmatrix}, \quad c = \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \\ 29/4 \end{bmatrix}.$$

The linearized problem is feasible only for $p = 4$ and 8; for $p \geq 16$, it is infeasible. This illustrates the remark made in the Introduction that some linearized problems may have feasible solutions even when the original problem (1) - (4) is actually infeasible.

Example 2 is the same as Example 1, except that constraints of type (4) are tightened so that they are active at the solution. Instead of (46), we take

$$a = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad d = \begin{bmatrix} .4 \\ .4 \end{bmatrix}.$$

The exact solution of this problem is $\varepsilon = \sqrt{2} - .4$, $z_1 = (-1 + i)\sqrt{2}/5$ and

$$z_2^I = \frac{317(\sqrt{2} - 1) - (431902 - 190320\sqrt{2})^{1/2}}{3000 - 1200\sqrt{2}} \doteq -.208846903$$

$$z_2^R = -\frac{\sqrt{2}}{10} + \left[\frac{1}{8} - \left(z_2^I - \frac{\sqrt{2}}{10}\right)^2\right]^{1/2} \doteq -.093336568$$

Tables 4 and 5 give, respectively, the solutions of the linearized primal and dual problems for $p = 4, 8, \ldots, 2048$. As in Example 1, the optimal ε for $p = 8$ is optimal also for all $p \geq 16$. The Richardson extrapolation of the last two z vectors in Table 4 can be performed as described in Example 1 above. In this case one extrapolation step gives

$$\begin{bmatrix} z_1^R \\ z_1^I \\ z_2^R \\ z_2^I \end{bmatrix} = \begin{bmatrix} -.282844 \\ +.282843 \\ +.0933341 \\ -.208848 \end{bmatrix}$$

which is an improvement of 2 to 3 significant digits over the best answer in Table 4.

p	4	8	16	32	64	128	256	512	1024	2048
z_1^R	-.400000	-.345442	-.293794	-.310700	-.296738	-.289786	-.286314	-.284562	-.283710	-.283277
z_1^I	.400000	.220244	.271891	.254985	.268948	.275899	.279372	.281123	.281975	.282409
z_2^R	.153553	-.0262742	-.0765711	-.0618571	-.0772609	-.0849613	-.0891368	-.0912435	-.0922757	-.0928049
z_2^I	-.153553	-.243431	-.217608	-.214390	-.211862	-.210582	-.209723	-.209290	-.209072	-.208960
ε	.600000	1.014214	1.014214	1.014214	1.014214	1.014214	1.014214	1.014214	1.014214	1.014214
Total Itera-tions	7	11	13	16	18	21	25	26	30	33

Table 4. Solutions of the linearized primal problem of Example 2.

p	4	8	16	32	64	128	256	512	1024	2048
basis names	1/2/1	1/2/8	1/2/15	1/2/29	1/2/57	1/2/113	1/2/225	1/2/449	1/2/897	1/2/1793
	1/2/4	1/3/6	1/3/12	1/3/22	1/3/43	1/3/85	1/3/169	1/3/337	1/3/674	1/3/1346
	2/1/3	2/1/4	2/1/7	2/1/13	2/1/25	2/1/49	2/1/97	1/3/338	2/1/385	2/1/769
	3/2/2	3/2/3	3/2/5	2/1/14	2/1/26	2/1/50	2/1/98	2/1/193	2/1/386	2/1/770
	3/2/3	3/2/4	3/2/6	3/2/10	3/2/19	3/2/36	3/2/71	3/2/141	3/2/280	3/2/558
basis values	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.
	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
	1.	1.	1.	1.	1.	1.	1.	0.	1.	1.
	0.	0.	0.	0.	0.	0.	0.	1.	0.	0.
	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.

Table 5. Optimal solutions of the dual problem of Example 2.

Example 3 is taken from [6] and is an unconstrained Chebyshev function approximation problem; that is, constraints (3) - (4) are absent. The 101 columns of the matrix $A \epsilon C^{3\times 101}$ are

$$A_j = \begin{bmatrix} 1 \\ \exp[i(j-1)\pi/400] \\ \exp[i2(j-1)\pi/400] \end{bmatrix} \quad , j = 1,2,\ldots,101$$

while the components of $f \epsilon C^{101}$ are

$$f_j = \exp\ [i3(j-1)\pi/400] \quad , j = 1,2,\ldots,101.$$

In other words, the complex valued function e^{i3x} is approximated by complex linear combinations of the three functions $\{1,e^{ix},e^{i2x}\}$ over 101 equispaced points on the x-interval $[0,\pi/4]$. As in Example 1 we insert bounds of type (4) to provide an initial dual basis; specifically, we take

$$a = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad , \quad d = \begin{bmatrix} 10 \\ 10 \\ 10 \end{bmatrix} .$$

These constraints are not active at the optimal solution. It is not necessary to insert artificial inequalities of type (3).

It can be verified that the explicit solution of Example 3 can be written

$z_1 = r_1 \exp\ (i\ 3\pi/8)$

$z_2 = r_2 \exp\ (i\ 5\pi/4)$

$z3 = r_3 \exp\ (i\ \pi/8)$

where

$$r_1 = a \doteq .96157056080646$$

$$r_2 = b-2(b-a^2)/(1-a^2) \doteq 2.8122548927058$$

$$r_3 = a(1-2b+a^2)/(1-a^2) \doteq 2.8477590650226$$

$$a = \lambda\cos(\pi/16)+(1-\lambda)\cos(\pi/8)$$

$$b = \lambda\cos(\pi/8)+(1-\lambda)\cos(\pi/4)$$

$$c = \lambda\cos(3\pi/16)+(1-\lambda)\cos(3\pi/8)$$

$$\lambda = (\sin(\pi/8))/(\sin(\pi/16)+\sin(\pi/8))$$

$$\varepsilon = (1-cr_1+ br_2 - ar_3)^{1/2} \doteq .014706309694449$$

Tables 6 and 7 give, respectively, the solutions of the linearized primal and dual problems for $p = 4, 8, \ldots, 1024$. The optimal value of ε for $p = 64$ seems to be the optimal ε value for all $p > 128$. The obvious "pairing" of the basic variables in Table 3 is explained by Theorem 5. Note also that the basis values are not altered, only permuted, for $p > 32$.

Richardson extrapolation can be applied to the last two z vectors in Table 6 just as in the two previous examples. In this case one extrapolation step gives

$$\begin{bmatrix} z_1^R \\ z_1^I \\ z_2^R \\ z_2^I \\ z_3^R \\ z_3^I \end{bmatrix} = \begin{bmatrix} .367978 \\ .888376 \\ -1.988564 \\ -1.988564 \\ 2.630986 \\ 1.089791 \end{bmatrix}$$

so that

$$|z_1| = .96157148$$

$$|z_2| = 2.81225418$$

$$|z_3| = 2.84775908 \ .$$

It is easy to verify that before extrapolation the case $p = 1024$ gives

$$|z_1| = .96236$$

$$|z_2| = 2.81376$$

$$|z_3| = 2.84855 \ .$$

Extrapolation in this case more than doubled the number of correct significant digits in the solution vector z.

Numerical computations for the above examples were performed on a DEC 10 at Stanford University.

p	4	8	16	32	64	128	256	512	1024
z_1^R	.362962	.378265	.377950	.377718	.372836	.370405	.369191	.368584	.368281
z_1^I	.953471	.913212	.912452	.911891	.900105	.894237	.891306	.889840	.889108
z_2^R	-2.018255	-2.026895	-2.024346	-2.022845	-2.005663	-1.997109	-1.992836	-1.990700	-1.989632
z_2^I	-2.119964	-2.026895	-2.024346	-2.022845	-2.005663	-1.997109	-1.992836	-1.990700	-1.989632
z_3^R	2.667545	2.654494	2.654624	2.654502	2.642716	2.636848	2.633917	2.632452	2.631719
z_3^I	1.154240	1.099528	1.099581	1.099531	1.094649	1.092218	1.091004	1.090397	1.090094
ε	.0122524	.0141560	.0145244	.0147063	.0147063	.0147063	.0147063	.0147063	.0147063
Total Iterations	11	20	25	33	36	39	42	45	48

Table 6. Solutions of the linearized primal problem of Example 3.

p	4	8	16	32	64	128	256	512	1024
basis names	1/1/1	1/1/8	1/1/14	1/1/28	1/1/56	1/1/112	1/1/224	1/1/448	1/1/896
	1/1/4	1/25/4	1/1/15	1/1/29	1/1/57	1/1/113	1/1/225	1/1/449	1/1/897
	1/24/2	1/28/5	1/26/7	1/26/13	1/26/26	1/26/53	1/26/104	1/26/209	1/26/417
	1/31/3	1/74/8	1/27/8	1/26/14	1/26/27	1/76/125	1/26/105	1/76/497	1/76/993
	1/78/1	1/77/1	1/75/16	1/76/32	1/76/63	1/76/126	1/76/249	1/76/498	1/76/994
	1/101/3	1/101/5	1/76/1	1/101/17	1/101/33	1/101/65	1/101/129	1/101/257	1/101/513
	1/101/4	1/101/6	1/101/9	1/101/18	1/101/34	1/101/66	1/101/130	1/101/258	1/101/514
basis values	.088039	.163234	.004365	.0	.0	.0	.0	.0	.0
	.082893	.244029	.160548	.168829	.168829	.168829	.168829	.168829	.168829
	.109350	.091325	.170912	.0	.0	.331171	.0	.331171	.331171
	.232546	.070677	.164573	.331171	.331171	.331171	.331171	.331170	.331170
	.302610	.263126	.173184	.331171	.331171	.0	.331171	.0	.0
	.158103	.157491	.162060	.168829	.168829	.168829	.168829	.168829	.168829
	.026457	.010118	.164358	.0	.0	.0	.0	.0	.0

Table 7. Optimal solutions of the dual problem of Example 3.

Another unconstrained Chebyshev function approximation problem given in [6] is moderately large and 100% dense. In this example the 501 columns of the matrix $A \in C^{44\times 501}$ are given by

$$A_j = \begin{bmatrix} \exp(i\,k_1\,x_j) \\ \exp(i\,k_2\,x_j) \\ \vdots \\ \exp(i\,k_{44}\,x_j) \end{bmatrix} - \exp(i\,k_{45}\,x_j) \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad , \; j = 1,2,\ldots,501$$

where $1 = k_1 < k_2 < \ldots < k_{44} < k_{45} = 49$ are the distinct integers between 1 and 49, excluding the integers $\{7,17,21,29\}$, and where $x_1 = u_0 + (j-1)(1- u_0)/250,\ j = 1,2,\ldots,501$ with $u_0 = .0538117$. The components of $f \in C^{501}$ are $f_j = \exp(i\,k_{45}\,x_j)$, $j = 1,\ldots,501$. This example lacks constraints of type (3) - (4). The linearized problem for $p = 16$ was solved on a DEC VAX 11/780 in 1350 simplex iterations. Total CPU time was 25 minutes and .7 million page faults were incurred. Only 80,000 words of storage were needed. In contrast, the algorithm proposed in [6] (which utilizes the algorithm [16] as a subroutine) solved the same problem on the same VAX in 1270 simplex iterations, requiring 179 minutes of CPU time and incurring 11 million page faults. Over 360,000 words of storage were needed. Both solutions were essentially identical, as expected. The difference in the number of simplex iterations is explained as follows. The algorithm [6] solves the full problem for $p = 16$, while the algorithm developed here solves the $p = 4$ problem and the $p = 8$ problem before solving the $p = 16$ problem. This indirect route to the full problem solution is less efficient in this example than solving the full problem immediately.

VI. Concluding remarks

Extrapolation was suggested as one method for extracting more accuracy from the linearized problem solutions. An alternative procedure can provide as much accuracy in the solution as desired, although with more work than is required by extrapolation. The optimal solutions of the linearized dual problems for sufficiently large p must identify the constraints active at the optimal solution of the original (i.e., $p = \infty$) problem (1) - (4). Deleting the inactive constraints from the $p = \infty$ problem yields an equality constrained optimiztion problem which the method of Lagrange multipliers seems well suited to solve. Lagrange's method gives rise to a nonlinear system of algebraic equations in the optimum value ε, the solution vector z, and the multipliers Λ. Iterative methods for the solution of this system can be started from a excellent initial point $(\varepsilon, z, \Lambda)$ provided by a linearized primal and dual problem solution. Safeguarded Newton-Raphson iteration should be a highly effective iterative method for solving this system, especially if advantage is taken of the system's special form (i.e., for Λ given, the vector z can be found by solving a system of linear equations). A possible limitation of this approach is that very large values of p might be necessary in order to identify the right $p = \infty$ active set. The examples of the previous section, however, indicate that the optimal active set is found for relatively small values of p. Specifically, in Examples 1, 2, and 3, the correct $p = \infty$ active sets (determined from the optimal dual basis names in Tables 3, 5, and 7) first appear when p is 8, 8, and 32, respectively.

Certain kinds of domain and range constraints can be adjoined to the constraints (8)-(10) with only minor extension of the algorithm proposed for solving (7) - (10). Let the matrix $H \in C^{n \times q}$, and the row vectors $e \in C^q$, $\phi \in R^q$, and $h \in R^q$ be given. Then the constraints

$$\mathrm{Re}\left((zH_j - e_j)\exp(-i\phi_j)\right) \leq h_j \ , \ j = 1,\ldots,q \tag{47}$$

are linear in z^R and z^I , and so can be added to the problem (7)-(10). It is easy to see that the constraints (9) and (10) are instances of (47). However, (47) can impose constraints not possible with (9) and (10). For instance, if q = 1, the constraint that the complex number $zH_1 - e_1$ must be in the right half of the complex plane is equivalent to

$$\mathrm{Re}\left((zH_1 - e_1)\exp(-i\pi)\right) \leq 0 \ . \tag{48}$$

Furthermore, if all the columns of the matrix H are identical to its first column H_1, then the number $zH_1 - e_1$ can be confined to any closed convex polygonal region (bounded or unbounded) in the complex plane by appropriate choices of ϕ, h, and q in (47). More generally, the constraints (47) can be thought of as confining each of the complex numbers $zH_j - e_j$ to different closed convex polygonal regions in the plane.

When complex function approximation on a discretized arc or domain boundary in the complex plane gives rise to the problem (1)-(4), then an implicit natural ordering of the columns of the matrix A exists. This ordering is inherited from the ordering of the discrete points along the arc. Therefore it is possible to devise clever strategies of both multiple and partial pricing in the simplex algorithm for solving the linearized problem in order to significantly reduce overall computation time. Effective partial pricing schemes would require far fewer evaluations of the vector-matrix products zA_j in (28) and yet not incur significant increases in the total number of simplex iteratios needed to reach optimality. Effective multiple pricing schemes would decrease the number of simplex iterations needed to reach optimality by increasing the change in ε in each simplex iteration step. Both multiple and partial pricing can be implemented simultaneously and might significantly improve computation time, especially when m and n are large.

It has been assumed throughout this report that the unknown vector z must lie in C^n. In some applications it is desirable to restrict z to R^n, while still retaining complex matrices A and B in original problem (1) - (4). It is easy to see that setting $z^I = 0 \in R^n$ in the linearized problem is equivalent to eliminating n of the 2n+1 rows of the Primal Problem constraints (22). All the techniques developed for the Dual Problem (23) - (24) simplify slightly when applied to the dual of this modified problem. Consequently the modified dual problem is smaller, easier to solve, and requires less storage than the unmodified dual.

Bibliography

1. K. Yosida, Functional Analysis, Second Edition, Springer-Verlag, 1968.

2. M.R. Osborne and G.A. Watson, "On the best linear Chebyshev approximation, "The Computer Journal, Vol. 10, 1967, pp. 172-177.

3. I. Barrodale and C. Phillips, "An improved algorithm for discrete Chebyshev linear approximation," Proc. of the Fourth Manitoba Conference on Numerical Math., B.L. Hartnell and H.C. Williams, Editors, Utilita Math. Pub. Co., 1974, pp. 177-190.

4. S.I. Zukhovitskiy and L.I. Avdeyeva, Linear and Convex Programming, W.B. Saunders Company, 1966. (Original Russian edition published in Moscow, 1964.)

5. K. Glashoff and K. Roleff, "A new method for Chebyshev approximation of complex-valued functions," Math. Comp., Vol. 36, No. 153, Jan. 1981, pp. 233-239.

6. R.L. Streit and A.H. Nuttall, "Linear Chebyshev complex function approximation and an application to beamforming," J. of the Acoustical Society of America, 72(1), July, 1982, pp. 181-190. (Also in Naval Underwater Systems Center Report 6403, 26 February 1981.)

7. R.L. Streit and A.H. Nuttall, "A note on the semi-infinite programming approach to complex approximation," Mathematics of Computation, to appear, April 1983.

8. D.G. Luenberger, Introduction to Linear and Nonlinear Programming, Addison-Wesley, 1973.

9. S.I. Gass, "Comments on the possibility of cycling with the simplex method," Letter to The Editor, Operations Research, Vol. 27, 848-852, 1979.

10. S.-A. Gustafson, "Nonlinear systems in semiinfinite programming," in Numerical Solution of Nonlinear Algebraic Systems, G.B. Byrnes and C.A. Hall (Editors), Academic Press, 1973, pp. 63-99.

11. S.-A. Gustafson and K. Kortanek, "Numerical treatment of a class of semiinfinite programming problems," Naval Research Log. Quart., vol. 20, 1973, pp. 477-504.

12. A. Ralston, A First Course in Numerical Analysis, McGraw-Hill, 1965.

13. D.C. Joyce, "Survey of extrapolation processes in numerical analysis", SIAM Review, Vol. 13, Oct. 1971, pp. 435-488.

14. K. Glashoff, "Duality Theory in Semi-Infinite Programming", in Semi-Infinite Programming, Lecture Notes in Control and Information Science, Vol. 15, 1979, pp. 1-16.

15. R.L. Streit, "Solution of large Hermitian eigenproblems on virtual and cache memory computers," ACM SIGNUM Newsletter, Vol. 16, 1981, pp. 6-7.

16. I. Barrodale and C. Phillips, "Solution of an overdetermined system of linear equations in the Chebyshev norm," ACM Algorithm 495, ACM Trans. on Math. Software, Vol. 1, No. 3, September, 1975, pp. 264-270.

INITIAL DISTRIBUTION LIST

Addressee	No. of Copies
ASN (RE&S)	1
OUSDR&E (Research & Advanced Technology)	2
Deputy USDR&E (Res & Adv Tech)	1
Deputy USDR&E (Dir Elect & Phys Sc)	1
OASN, Spec Dep for Adv Concept	1
Principal Dep Assist Secretary (Research)	1
OASN, Dep Assist Secretary (Res & Applied Space Tech)	1
OASN, Director, Submarine & ASW Diagrams	1
ONR, ONR-100, -102, -200, -220, -400, -410, -414, -420, -422, -425, -425AC, -430	12
CNO, OP-098	1
CNM, MAT-05, SP-20, ASW-13	3
NRL	1
NORDA	1
OCEANAV	1
FNOC	1
NAVOCEANO, Code 02	1
NAVELECSYSCOM, ELEX 03	1
NAVSEASYSCOM, SEA-614	1
NAVAIRDEVCEN, Warminster	1
NAVAIRDEVCEN, Key West	1
NOSC	1
NAVWPNSCEN	1
NCSC	1
NAVCIVENGRLAB	1
NAVSWC	1
NAVSURFWPNCEN	1
NWTNSRDC ANNA	1
NWTNSRDC BETH	1
NAVPGSCOL	1
APL/UW, SEATTLE	1
ARL/PENN STATE, STATE COLLEGE	1
DTIC	1
DARPA	1
NOAA/ERL	1
NATIONAL RESEARCH COUNCIL	1
WOODS HOLE OCEANOGRAPHIC INSTITUTION	1
ENGINEERING SOCIETIES LIB, UNITED ENGRG CTR	1
NATIONAL INSTITUTE OF HEALTH	1
ARL, UNIV OF TEXAS	1
MARINE PHYSICAL LAB, SCRIPPS	1
UNIVERSITY OF CALIFORNIA, SAN DIEGO	1
NAVSURWEACTR	1
INTERIMS INC.	1
MAR INC.	1
B-K DYN INC.	1
BBN	1
EWASCTRI (T. Russell)	1
HYDROINC (D. Clark)	1
SUMRESCR (M. Henry)	1

INITIAL DISTRIBUTION LIST (Cont'd)

Addressee	No. of Copies
ANALTECH INC. (G. Bourgond)	1
ANALTECHNS (T. Dziedzic)	1
GENPHYSORP (M. Baver)	1
EDOCORP (J. Vincenzo)	1
OPERRES INC. (Dr. V. P. Simmons)	1
TRA CORP. (J. Wilkinson)	1

DATE
ILME